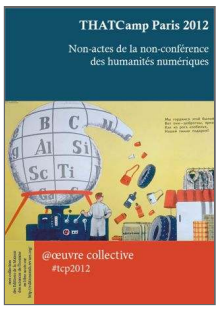

THATCamp Paris 2012
Non-actes de la non-conférence des humanités numériques

Éditions de la Maison des sciences de l'homme

Utilisons RDFa pour nos corpus

Proposé par Stéphane Pouyllau

Collectif

DOI : 10.4000/books.editionsmsh.379
Éditeur : Éditions de la Maison des sciences de l'homme
Lieu d'édition : Paris
Année d'édition : 2012
Date de mise en ligne : 1 octobre 2012
Collection : La Non-Collection
ISBN électronique : 9782735115273

Utilisons RDFa pour nos corpus

Proposé par Stéphane Pouyllau

Collectif

Introduction de Stéphane Pouyllau

1 Il s'agit avec cet atelier d'envisager la question de la publication sur le Web à l'aide du Web de données, de la portabilité du RDFa dans le monde des SHS. C'est-à-dire, de voir comment préparer l'exposition de données numériques complexes dans le Web de données, selon les principes du linked data, et réfléchir ensemble à la manière dont nous pourrions dépasser la phrase habituelle qui consiste à dire « Je vais publier sur le Web ».

2 On peut comprendre l'utilisation du RDFa comme quelque chose qui nous offre la capacité, lorsque on fait du traitement sur des données structurées depuis le Web, de faire des traitements augmentés, en reliant l'information à des informations contenues dans des documents distants, ceci en rendant explicite pour des machines, à la fois la structure, mais aussi le contenu sémantique des documents.

3 L'idée de cet atelier est de voir comment la mécanique du RDF paraît bien s'appliquer au type de travail des chercheurs en sciences humaines et sociales, au sens où il s'agit de produire de l'information, de la publier et de la rendre explicite. Cela pourrait être en quelque sorte la marche supérieure vers une meilleure interopérabilité des informations et donc des données après l'OAI[1] avec l'utilisation de Dublin Core[2] : jusqu'à présent on était obligé de coller un petit programme aux bases de données pour exprimer les données, toute modification de la structure de la base de données impliquait donc la réécriture de l'API, RDFa est peut-être la seconde marche vers l'interopérabilité, c'est-à-dire le moment où les chercheurs vont reprendre leurs données et rendre explicite la structuration de toute cette information afin de la publier dans le Web de données.

4 Deux grandes notions doivent être posées en introduction afin de mieux débattre, celles des syntaxes RDF et RDFa, et le concept de Web de données et de linked data. Comment faire prendre conscience aux chercheurs de l'importance de la structuration, y compris dans le Web pour permettre aux moteurs de recherche de mieux retrouver l'information

et les données (en passant par des métadonnées exprimées avec plus de richesse sémantique).

L'utilisation des syntaxes RDF et RDFa

5 L'utilisation du RDFa est une manière d'exprimer du RDF dans une page web. La norme décrit la manière d'enchâsser de la sémantique à l'aide d'attributs ajoutés dans les balises HTML qui vont pointer vers des URI[3]. On peut dire que RDFa est une sorte de « micro-formats ». On peut tout à la fois, à l'aide de ce mécanisme, exprimer la structure des pages ou décrire explicitement que telle information désigne le lieu de naissance d'un auteur, la géolocalisation d'un lieu vis-à-vis d'un référentiel géographie présent dans *linked data* (tel que GeoNames.org), etc.

Le RDFa selon le W3C

6 Pour décrire cette information, on a recours à des modèles qu'on appelle des ontologies. Divers sites référencent de telles ontologies. Citons par exemple Schemapedia ou encore LOV (Link open vocabulary), un site développé initialement par la société Mondeca et rendu public depuis dans le cadre du projet Datalift. Vouloir rendre explicite de l'information dans une page donnée, et la manière de le faire, relèvent de choix scientifiques. Plusieurs outils permettent ensuite de détecter la présence de RDFa dans les pages web. Toutefois, il n'y a pas encore beaucoup d'outils d'un niveau suffisant pour l'exploitation de cette structuration.

7 Au cours de la discussion, Lou Burnard explique qu'il hésiterait à dire qu'il s'agit d'une structuration, il s'agit plutôt pour lui de « labeliser » (au sens d'étiqueter) telle ou telle partie du texte[4]. L'attribut définit tel élément et explique où aller chercher pour obtenir une définition. En revanche, il fait remarquer que cela ne permet pas de savoir ce qu'on peut faire avec. Avec la TEI, on est dans le cadre d'un balisage structurel et informationnel dans un contexte XML. Tandis que dans le cas de l'utilisation du RDF ou de RDFa, on n'est pas forcément toujours dans un environnement XML, et par ailleurs on va pouvoir aller chercher de l'information qui peut être contenue dans des réservoirs qui ne sont pas nécessairement contenus dans la ressource elle-même, mais ailleurs sur le Web. La TEI offre une sémantique, elle est exprimée selon un modèle d'arbre avec XML, le modèle de graphe RDF avec ses relations entre des entités (sujet-prédicat-objet) permet la mise en relation de sources. Ce dernier n'est pas contradictoire avec TEI. On peut dire, en fait, qu'avec RDF on fournit des méta-informations. Si par exemple on désigne un auteur, on informe sur l'endroit où trouver de l'information sur cet auteur même si les traitements ne sont pas définis par l'encodage[5].

8 L'un des participants souligne la difficulté qu'il y a à identifier des ontologies adaptées ou efficaces dans un contexte de recherche en sciences humaines et sociales. Les répertoires d'ontologies disponibles ne permettent pas toujours d'identifier clairement la portée ou la précision de ces ontologies et leur applicabilité. Les modèles les plus connus comme FOAF (Friend of a Friend), peuvent être d'une portée limité pour ce qui concerne, par exemple, la description des personnes historiques. Cela constitue sans doute un frein pour la généralisation du recours à RDFa. Toutefois, le modèle conceptuel de référence CIDOC (CIDOC-CRM) offre déjà une ontologie consistante pour les sciences historiques et l'histoire de l'art. Le risque serait de tomber dans l'excès inverse en se lançant pendant 15 ans dans la mise au point d'une ontologie parfaite, en ne faisant rien en attendant.

9 Le RDFa permet par exemple d'exprimer une ontologie dans un texte. Un participant fait état de ses expérimentations dans des textes historiques encodés en XML-TEI pour identifier des personnes travaillant dans le même endroit. Le RDFa n'est au fond que le véhicule de ce que fait une ontologie, cela nécessite au préalable de définir (ou de choisir l'ontologie). Un premier travail consiste pour l'historien à savoir quelles sont les ontologies qui peuvent lui permettre d'exprimer toute la richesse de ce qu'il veut dire. Il convient donc au préalable de se mettre d'accord sur la bonne ontologie à utiliser pour exprimer un travail sur un texte. Comment par ailleurs exprimé toute la richesse de l'histoire ; et l'incertitude ?

Identifier les ontologies utiles aux SHS

10 La discussion conduit les participants à envisager la mise sur pied d'un site web répertoriant les ontologies disponibles pour les chercheurs en sciences humaines et sociales (historiens, historiens de l'art, sociologues, etc.), en soulignant leurs applicabilité, leur degré de finesse, et en indiquant par qui elles sont maintenues. Un tel site pourrait également, outre des guides pratiques, présenter des exemples de réalisations. Certains participants alertent sur le risque de recloisonnement des disciplines avec des répertoires spécialisés. Même si le modèle conceptuel de référence CIDOC (CIDOC-CRM) a été conçu dans le domaine de l'histoire de l'art, certains éléments qui le composent peuvent être utiles pour d'autres disciplines par exemple. Il y a déjà deux ans, un groupe de travail avait ainsi travaillé sur l'expression de la TEI avec le modèle conceptuel de référence CIDOC (CIDOC-CRM).

11 Dans le cadre du projet Athena, et en lien avec Europeana, un travail de recensement avait déjà permis d'identifier récemment un très grand nombre de vocabulaires spécialisés utilisés dans le monde des musées en Europe. Plus de 150 vocabulaires avaient été identifiés, la difficulté consistait à les classer d'après des critères efficaces, par domaines, par disponibilités en *linked data* (ce qui n'avait pas été pensé à l'époque), etc. Le projet Linked heritage qui lui fait suite, permettra de caractériser plus précisément les liens de *mapping* avec une plate-forme fournissant toutes les relations typées, avec des indications sur la finesse des relations, exprimées en SKOS (Simple Knowledge Organization System). Une telle plate-forme devrait rendre possible le jeu entre des silos très spécialisés et devrait éviter toute velléité de produire, *de nuovo*, une ontologie mondiale hyperspécialisée, ce qui serait tout simplement impossible.

12 Si la première marche dans la direction de l'interopérabilité des données scientifiques en sciences humaines et sociale a été la généralisation de l'emploi du Dublin Core via le

protocole OAI-PMH, on peut se féliciter déjà de l'emploi de ces six verbes OAI-PMH partout dans les SHS et au niveau mondial. Bien sûr, le Dublin Core simple lissait par trop l'expression de l'information, ainsi certains ont pu avoir recours au Dublin Core étendu (DC Terms). Si l'utilisation du RDFa constitue une deuxième marche vers l'interopérabilité des données scientifiques, le but est peut-être d'abord d'amener les gens à exprimer ce type de structures avec ces modèles. Dans tous les cas, cela passe par des travaux communs et interdisciplinaires pour ne pas retomber dans le travers de l'ultra-spécialisation des ontologies.

Automatiser la sémantisation des corpus

13 Pour les corpus existants, deux cas de figure peuvent être envisagés : amener les chercheurs à revenir sur leurs documents pour faire de l'enrichissement sémantique, ce qui peut être une solution lourde, ou bien essayer de produire du balisage automatique. Dans tous les cas, il convient d'être attentif au gain de l'investissement en termes d'utilisation.

14 Stéphane Pouyllau ne croit pas trop à la sémantisation automatique, une sémantisation semi-automatique serait en revanche envisageable car il faudra toujours faire des choix. Un participant mentionne, à titre d'exemple, la plate-forme d'annotation sémantique développée par l'Institut de recherche et d'innovation du centre Pompidou (IRI) pour produire une exploration sémantique du portail Histoire des Arts[6]. Même dans un projet comme celui de l'extraction des données structurées des pages de Wikipédia en français, DBpedia en français, à partir des infobox de l'encyclopédie en ligne, les disparités de présentation nécessitent des ajustements manuels.

15 Il est peut être aussi souhaitable que les chercheurs soient directement confrontés à l'encodage de ces informations car cela pourrait présenter pour eux un intérêt en terme de questionnements scientifiques. Un participant fait remarquer qu'en 2006 des chercheurs en bio-informatique ont déjà été confrontés à des questions comparables et qu'il pourrait être profitable de prendre contact avec eux[7].

Inciter à l'utilisation d'une structuration de l'information exprimée à l'aide de la syntaxe RDFa

16 L'utilisation d'une structuration/étiquetage (labelisation) à l'aide de la syntaxe RDFa dépend très directement des possibilités d'utilisation. Le moteur de recherche Isidore est actuellement l'un des seuls capables de récupérer de l'information disponible sur les pages web en RDFa et de la structurer. Le fait qu'Isidore moissonne les métadonnées est déjà très incitatif[8]. Il a été tenu compte de cette possibilité lors de la conception du moteur dans le cadre d'un travail avec l'Agence bibliographique de l'enseignement supérieur (Abes) qui produisait à l'époque l'application Calames pour le catalogue en ligne des archives et des manuscrits de l'enseignement supérieur et où une structuration RDFa avait été enchâssée dans les pages. Cela avait été réalisé à titre expérimental, mais on a pu constater que le connecteur fonctionnait. Depuis, de nombreux webmasters s'adressent à Isidore pour signifier que le simple moissonnage RSS ne leur suffit plus en voulant proposer des informations plus précises exprimées à l'aide de RDFa.

17 Il est probable que les webmasters se tournent de plus en plus vers RDFa car les moteurs de recherche comme Google s'y intéressent[9] et que cela peut avoir des répercussions sur les résultats de la recherche. De fait, les possibilités techniques ouvertes par les technologies du web 3.0 (sémantique + LOD Linked Open Data) devraient produire un appel d'air et un cycle d'incitations vertueuses du côté des webmasters.

18 Afin d'inciter les chercheurs à l'emploi du RDFa pour la structuration (ou l'étiquetage) de l'information en sciences humaines et sociales publiée sur le Web, on souligne la nécessité d'expliquer les possibilités offertes par une telle description et ses avantages. Il s'agit non seulement de leur donner une idée sur la manière dont cela pourra leur être utile, mais aussi de leur expliquer concrètement ce qu'il faut faire. Aussi, les participants de l'atelier reviennent sur l'idée d'un site web qui recense les ontologies disponibles en sciences humaines et sociales, mais qui fasse peut-être plus : la présentation d'exemples de *mashups* pourrait inciter, par la visualisation de ce que l'on peut faire avec des données, à l'utilisation de cette technologie. Outre des démonstrateurs du potentiel de cet étiquetage, des guides pratiques seraient utiles. La rhétorique consisterait à montrer d'abord que c'est sexy, et ensuite à prouver que cela n'est pas si compliqué à mettre en œuvre.

19 Un participant fait aussi remarqué que l'on obtiendra sans doute une adhésion des chercheurs en sciences humaines et sociales et leur participation à partir du moment où l'on pourra arriver à exprimer une fiabilité dans l'information ou à exprimer des notions de débats. C'est sans doute à ce niveau que se joue l'avenir de l'utilisation de RDFa dans le monde des SHS. L'intérêt principal d'une structuration de l'information avec RDFa est sa capacité à introduire une information sur la fiabilité de l'encodage. On a souvent à faire en sciences humaines au problème de l'incertitude. Peut-être que le modèle RDF nous permettra de traiter l'incertitude. Il serait peut-être opportun de créer un centre de ressources pour traiter de cette question.

20 S'il s'agit de faciliter l'emploi de RDFa dans le cadre de projets en XML-TEI, il faut fournir au chercheur des propositions d'attributs qu'il puisse utiliser dans son travail. Lorsqu'un chercheur veut bien ajouter des descripteurs, où peut-il trouver des propositions, un mapping des attributs, etc. dont Isidore va se servir ? Comment traduire en RDFa de la sémantique contenu dans les attributs des balises TEI dans le HTML lorsque l'on fait la transformation ? Il manque une documentation et des exemples pour trouver cette information[10].

21 Afin qu'il soit bien compris qu'il n'y a pas de contradiction entre « la communauté RDF » et « la communauté TEI », il convient de bien expliquer comment exprimer du RDF dans des documents TEI. Le trésor de guerre des chercheurs se sont les documents XML-TEI sur lesquels ils travaillent souvent depuis plusieurs années. Ils pensent sans doute à raison qu'il serait dangereux de capitaliser uniquement du RDF dans du HTML qui pour eux constitue un format de sortie[[note : L'exemple proposé dans l'atelier, une page HTML avec du RDFa, n'est qu'un exemple et doit être vu comme une « sortie » en ligne, comme une « édition » HTML et pas comme conteneur pérenne de données et d'information.]]. Une des manière d'inciter la communauté des humanités numériques à employer cette structuration de l'information est de leur fournir des moyens sûrs de l'exprimer dans des documents XML-TEI et de leur donner des moyens de l'exporter de manière fiable. On souhaite pérenniser le travail, et le faire une seule fois, en ce sens la notion de workflow est à prendre en compte.

22 D'après Stéphane Pouyllau, le problème n'est pas, alors que l'on passe déjà beaucoup de temps à encoder en TEI, qu'aujourd'hui des informaticiens viennent nous demander d'exprimer les choses en RDF. Quelque soit le langage balisé, on est capable, selon lui, de reconstituer des formats de sortie. De la même manière que de fabriquer des bases de données relationnelles n'empêche pas d'exprimer les choses en RDF, il n'y a pas de problèmes avec la TEI. Afin d'amener une communauté des humanités numériques très largement XML-TEI à inscrire des métadonnées RDFa dans le HTML, il faudrait peut-être alors que la communauté RDF et TEI rédige ensemble une spécification simple exprimant de la TEI en RDF, et proposer des *web services* de génération du HTML/RDFa correspondant pour les documents TEI conformes à cette spécification. Cela existe sans doute.

23 Pour ceux qui ne sont pas encore passés au XML-TEI, peut être convient-il également de développer des logiciels performants pour faire de l'annotation RDF. Par ailleurs, il convient de trouver la bonne ontologie et savoir si l'information que l'on décrie existe quelque part. Un des problèmes aujourd'hui est qu'il n'y a pas encore de moteurs de recherche pour savoir si l'information est déjà décrite quelque part. Enfin, il n'existe pas vraiment d'outils accessibles à des non-informaticiens qui permettent une réelle démocratisation du RDFa. L'exploitation des informations en RDFa nécessite souvent des développements à façon.

24 L'atelier débouche sur la proposition d'un site web qui :

25 - référencerait et documenterait les ontologies disponibles en LOD pouvant intéresser les chercheurs en sciences humaines et sociales, c'est à dire classées, décrites, selon les besoins des disciplines ;

26 - comporterait des tutoriels d'utilisation du RDFa ;

27 - fournirait une indication concernant les machines en mesure d'exploiter dès à présent ou à terme les attributs RDFa ;

28 - offrirait une documentation sur les attributs exploitables par les machines « RDFa-compliant » ;

29 - présenterait des démonstrations sur les utilisations possibles et des *mashups* mettant en évidence les possibilités imprévues ;

30 - rassemblerait des guides « pas par pas » faisant sentir que le coût d'entrée n'est pas si élevé, ces guides se déclinant en fonction des produits finis que souhaiterait obtenir le chercheur.

31 Il pourrait subsidiairement :

32 - fournir des applications automatisant un peu l'expression en RDFa des attributs dans le document XML-TEI.

BIBLIOGRAPHIE

Webographie

Stéphane Pouyllau, « Construire le web de données pour les données de la recherche en SHS : comment utiliser RDFa ? », *sp.Blog*, 30 août 2010, http ://blog.stephanepouyllau.org/401, consulté le 27 septembre 2012.

Got, « Comprendre RDFa en 5 minutes », *Les petites cases*, 6 décembre, 2008, http ://www.lespetitescases.net/comprendre-rdf-en-moins-de-5-minutes, consulté le 27 septembre 2012.

Got, « Comprendre RDFa en 5 minutes », *Les petites cases*, 6 décembre, 2008, http ://www.lespetitescases.net/comprendre-rdf-en-moins-de-5-minutes, consulté le 27 septembre 2012.

Got, « RDFaiser votre blog, 1ère partie : la théorie », *Les petites cases*, 24 février, 2008, http ://www.lespetitescases.net/comprendre-rdf-en-moins-de-5-minutes, consulté le 27 septembre 2012.

Got, « RDFaiser votre blog, 2ème partie : la pratique », *Les petites cases*, 24 février, 2008, http ://www.lespetitescases.net/rdfaiser-votre-blog-2-la-pratique, consulté le 27 septembre 2012.

Got, « RDFaiser votre blog, 3ème partie : l'exploitation », *Les petites cases*, 24 mai, 2008, http ://www.lespetitescases.net/rdfaiser-votre-blog-3-exploitation, consulté le 27 septembre 2012.

Adida, Ben and Mark Birbeck, éditeurs, « RDFa 1.1 Primer, Rich Structured Data Markup for Web Documents », W3C Working Group Note, 7 juin 2012, http ://www.w3.org/TR/xhtml-rdfa-primer, consulté le 27 septembre 2012.

Dan Brickley, » Introducing FOAF », *FOAF Project*, 2000, http ://www.foaf-project.org/original-intro, consulté le 27 septembre 2012.

The Athena WP4 wiki, http://www.athenaeurope.org/athenawiki, consulté le 27 septembre 2012.

Got, « XML vs RDF : logique structurelle contre logique des données », *Les petites cases*, 29 août, 2010, http ://www.lespetitescases.net/xml-vs-rdf, consulté le 27 septembre 2012.

Normes et standards

Standards RDF sur le site du W3C, http ://www.w3.org/standards/techs/rdf, consulté le 27 septembre 2012.

Standards RDFa sur le site du W3C, http ://www.w3.org/standards/techs/rdfa, consulté le 27 septembre 2012.

SKOS Simple Knowledge Organization System, http ://www.w3.org/2004/02/skos, consulté le 27 septembre 2012.

Lou Burnard et Bauman Syd éditeurs, « P5 : Guidelines for Electronic Text Encoding and Interchange. », *Text Encoding Initiative*, 2007, http ://www.tei-c.org/release/doc/tei-p5-doc/en/html, consulté le 27 septembre 2012.

ICOM, *CIDOC Conceptual Reference Model*, http://www.cidoc-crm.org, consulté le 27 septembre 2012.

NOTES

1. Voir : http://fr.wikipedia.org/wiki/Dublin_Core.
2. Voir : http://fr.wikipedia.org/wiki/Dublin_C

3. Voir : http://fr.wikipedia.org/wiki/Uniform_Resource_Identifier
4. Cette question terminologique est revenue de manière récurrente au cours de la discussion. Nous employons dans ce compte-rendu les termes « structuration de l'information » qui nous semble moins problématique que le mot structuration employé seul, et « étiquetage » à la place de labelisation qui est un anglicisme
5. Sur la différence de logique entre XML-TEI et RDF, voir ce billet du blog Les petites cases et les commentaires qu'il a suscités.
6. Voir la démonstration du service web sur le site de l'IRI.
7. L'équipe mentionnée au cours de l'atelier s'intitule Modélisation conceptuelle des connaissances médicales, unité INSERM U936 de l'université Rennes-1.
8. Voir la documentation de la plate-forme de recherche Isidore.
9. Voir à ce sujet l'initiative propre à Google, schema.org mais aussi les « rich snippets » de google.
10. Le Centre de ressources numériques (MEET) pourrait peut-être proposer cela ?

RÉSUMÉS

L'atelier consiste à évoquer la qualité des métadonnées et des données de nos corpus numériques à la suite de la non-conférence de Paul Bertrand. Comment dépasser le plus petit dénominateur d'aujourd'hui, souvent le Dublin Core, pour aller plus loin ? En discutant autour du modèle RDF et plus particulièrement du RDFa pour le web (HTML), il s'agit d'entrevoir ce qu'il est possible de faire en matière de guide de bonnes pratiques, techniques, vecteurs, comment mieux « pousser » des données, dans Isidore par exemple.